Data Mining Project: Part 1
Data Analysis and Preprocessing

Real-world data is often noisy, large, and comes from diverse sources, making the first step in data mining an exploratory data analysis. This involves examining the attributes, their values, distributions, and identifying issues like outliers or missing data. Following this, data preprocessing is crucial to ensure quality. It includes cleaning to remove noise, integrating data from various sources, reducing data size, and transforming it through normalization to enhance the performance of mining algorithms. These preprocessing steps can be applied together for better results.

So, in this first part of the project, you are asked to familiarize yourself with your data and extract as much information as possible from it. As the Climate dataset covers the entire world, you are asked to select the Wilaya of Tizi-ouzou to work on for the entire project. The work requires the design and implementation of an application that should allows at least:

A. Data manipulation:
   a. Import, visualize, and save the contents of a dataset.
   b. Provide a global description of the dataset.
   c. Update/Delete an instance or value of the dataset.

B. Analysis of the characteristics of the dataset attributes:
   a. For each attribute:
      i. Calculate measures of central tendency and deduce symmetries.
      ii. Calculate measures of dispersion and deduce outliers.
      iii. Calculate the amount of missing values and unique values.
      iv. Construct boxplots and display outliers.
      v. Construct histograms and visualize the data distribution.
   b. Construct and display scatter plots of the data and deduce correlations.

Thus, you are asked to preprocess the dataset. It is imperative to use the analysis carried out in the 1st step to guide you in this 2nd step. The dataset obtained after applying all the necessary treatments must be 100% functional and as optimal as possible for the next steps. Add to your interface the following functionalities:

C. Data reduction through aggregation by seasons.
D. Data integration: merges data from multiple sources into a single coherent dataset.
E. Multiple choices of handling outliers and missing values.
F. Data normalization: Min-Max / z-score methods.
G. Data reduction via discretization of continuous data: Equal Frequency / Amplitude.
H. Data reduction (elimination of redundancies) horizontal / vertical.

IMPORTANT! For the final dataset, you have to choose the right preprocessing steps to apply and in what order. This is of utmost importance and your choices must be justified.

Deadline: Tuesday, November 11, 2024 at 1p.m.

Have fun !

HOUACINE Naila Aziza : n.a.houacine@gmail.com